Learning Pairwise Feature Dissimilarities for Person Re-Identification

Niki Martinel
University of Udine
Udine, Italy
niki.martinel@uniud.it

Christian Micheloni
University of Udine
Udine, Italy
christian.micheloni@uniud.it

Claudio Piciarelli
University of Udine
Udine, Italy
claudio.piciarelli@uniud.it

***Abstract*—This paper deals with person re-identification in a multi-camera scenario with non-overlapping fields of view. Signature based matching has been the dominant choice for state-of-the-art person re-identification across multiple non-overlapping cameras. In contrast we propose a novel approach that exploits pairwise dissimilarities between feature vectors to perform the re-identification in a supervised learning framework. To achieve the proposed objective we address the person re-identification problem as follows: i) we extract multiple features from two persons images and compare them using standard distance metrics. This gives rise to what we called distance feature vector; ii) we learn the set of positive and negative distance feature vectors and perform the re-identification by classifying the test distance feature vectors. We evaluate our approach on two publicly available benchmark datasets and we compare it with state-of-the-art methods for person re-identification.**

I. Introduction

Recent achievements of image sensing technologies have boosted video analytic systems for wide area surveillance. Though the sensor devices are becoming cheaper, monitoring a wide area by deploying a large number of cameras is still a challenging task. As a matter of fact, in wide area surveillance systems, not all the zones are usually covered by sensors. This opens up to the critical issues of uncovered areas, named "blind gaps". As a result of the "blind-gaps", large variation of viewpoints, illumination conditions, scales, etc., the task of re-identifying a person moving across such uncovered areas is a challenge to the community.

To deal with these issues, during the last years much effort has been made to design robust features that can be used to describe and match a specific person across different cameras [6], [12], [11]. Different approaches have been proposed to find linear and non-linear [14], [9] transformation functions between appearance features among pairs of cameras. The problem of target re-identification has also been addressed by finding the best distance metric [8], [18] that can be used to match features across non-overlapping cameras.

Despite this, target re-identification in a non-overlapping multi-camera scenario is still an open issue due to the challenging issues of pose variations, illumination changes, that introduce unknown transformations of the features between cameras. Motivated by the recent success of metric learning methods and feature transformation approaches we propose a novel target re-identification approach to address these challenges. The core novelty of this work is a method that aims to model not the way features are transformed across camera, but to advantage of invariant features and proper distance metrics to model how the distances between such features are transformed across cameras. To achieve this goal we extract the feature vectors from a pair of targets viewed in different cameras, then we compute the distance between such features and use the distances to form the distance feature vector (DFV). The DFVs from the same person form the set of positive samples, while the DFVs from different persons form the negative set. Using the positive and negative DFVs we re-identify persons in a supervised classification framework.

Positive and negative samples are classified by means of a DFV applied on a pair of cameras. The proposed solution can be easily included in a distribute framework like that we previously proposed in [13].

To validate the proposed method we compare the performance of our approach to state-of-the-art methods for person re-identification using two publicly available benchmark datasets.

II. Related work

In the past different approaches have been proposed by the community to deal with the problem of person re-identification across non-overlapping cameras: i) methods that use invariant appearance features, ii) methods that capture the transformation of features across camera pairs, and iii) methods that learn the optimal distance metric between appearance features.

Invariant feature methods are the most commonly explored approaches for person re-identification. Particular interest has been focused on finding the best set of features [11] that can be exploited to match persons across cameras. In [4], [12], [2] multiple local and global features were used to create an invariant signature using multiple image frames. In [15], frames were used to built a collaborative representation that best approximates the query frames. In [10], the distribution of color features projected in the log-chromaticity space was described using the shape context descriptor. In [17] an unsupervised framework was proposed to extract distinctive features. A patch matching method was used together with adjacency constraint to tackle viewpoint changes and pose variations.

Transformation methods were been explored in [9] to capture the transformation across non-overlapping cameras in a tracking scenario. Similarly, the problem of capturing the non-linear transformation between features was addressed in [14]. In [1] the implicit transformation function of features was

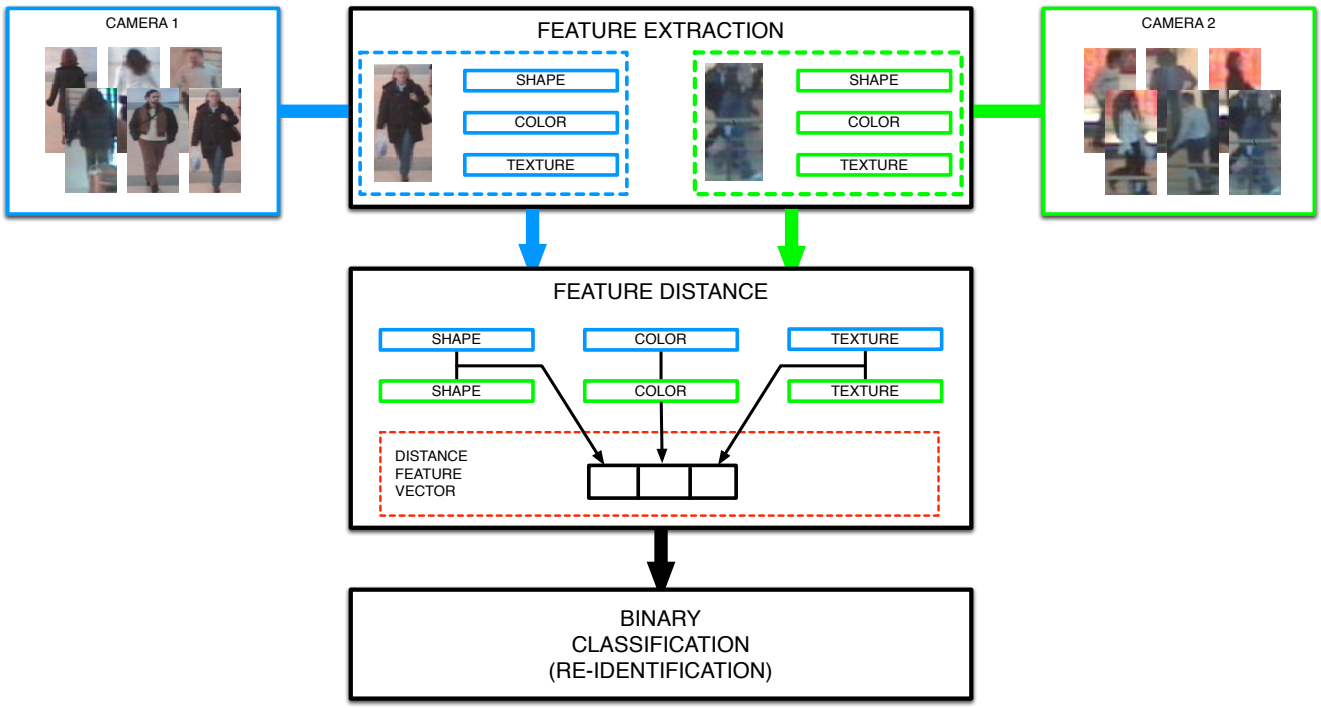

Fig. 1. An overview of our person re-identification approach. From each given image, we extract shape, color and texture features, then we compute the pairwise distances between the feature vectors extracted for targets viewed by different cameras. The computed distances form the DFV. The DFV from a pair of images of the same person is a positive sample, while the DFV from a pair of images of different persons is a negative sample. The DFVs are used to train a binary classifier. The trained classifier is used to re-identify targets by classifying test DFV.

learned by concatenating appearance feature vectors of persons viewed by different cameras.

Distance methods learn the best metric between appearance features of the same person across camera pairs. In [8] the Largest Margin Nearest Neighbor technique was exploited to learn a Mahalanobis metric using pairs of labeled samples from different cameras. In [5] the Large Margin Nearest Neighbor with Rejection method was proposed. In [16] the re-identification problem was formulated as a local distance comparison problem introducing an energy-based loss function that measures the similarity between appearance instances.

III. Our approach

An overview of our approach is shown in Figure 1. Given a pair of images from non-overlapping cameras, we extract multiple local and global features (Section III-A) and we compute pairwise distances between them (Section III-B). The computed distances form the DFV. To use this feature for classification we train a binary classifier to classify new examples.

A. Feature extraction

Here we describe the feature extraction methods used to build a discriminative representation of the image of a person.

Motivation. The task of re-identifying targets across camera pairs is challenging because of the issues of pose variation, illumination and color changes. State-of-the-art methods for person re-identification have successfully explored different appearance features [11] to tackle these challenges. Inspired by that, to obtain a robust feature representation of an image across cameras, we considered, shape, color and texture features invariant to the stated issues.

Shape. To capture the shape of a given person we used the Pyramid Histogram of Oriented Gradients (PHOG) feature. The PHOG feature is computed exploiting the spatial pyramid technique. Let $l = 0, \cdots, L$ be the level of the spatial pyramid, and 4^l the number of cells in which the image is divided at each level l. The PHOG feature Φ is the concatenation of the HOG computed at the different levels and for different cells of the spatial pyramid. The final PHOG feature vector is of size $b \sum_{l=0}^{L} 4^l$, where b is the number of bins used to compute the HOG features.

Color. Color histogram features are the most widely used features to describe a person image. All state-of-the-art person re-identification methods use color features relying on the assumption that persons do not change their clothes as they move between camera Fields-of-view. According to that, we extract six different color histogram features from each given image. We consider that most of the persons wear different clothes for the upper and lower body part, so, before computing the color features we detect the three salient body parts (i.e., head, torso and legs) using a derivation of the approach proposed in [7]. We discard the head region from the feature computation since it generally contains few and not informative pixels. To achieve illumination invariant properties we equalize the histograms of the two regions and project them into the Lab color space. Then, we extract a histograms for each color channel c for both the torso and legs regions. The histograms for the two regions are denoted $\Upsilon_T \in \mathbb{R}^{n_c}$ and $\Upsilon_L \in \mathbb{R}^{n_c}$, respectively.

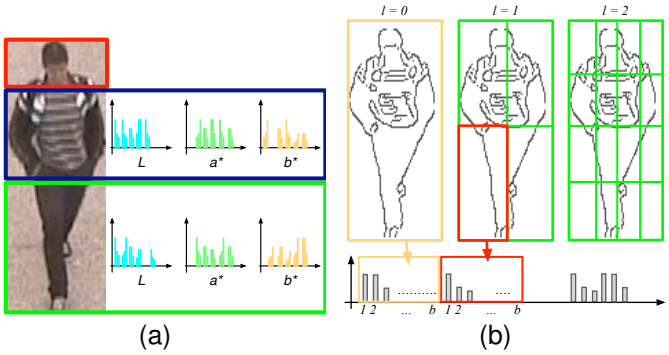

Fig. 2. Color and shape features. (a) Color histogram features extracted from the torso and legs body parts. (b) PHOG features extracted from the whole body at three different levels of the spatial image pyramid ($L = 2$).

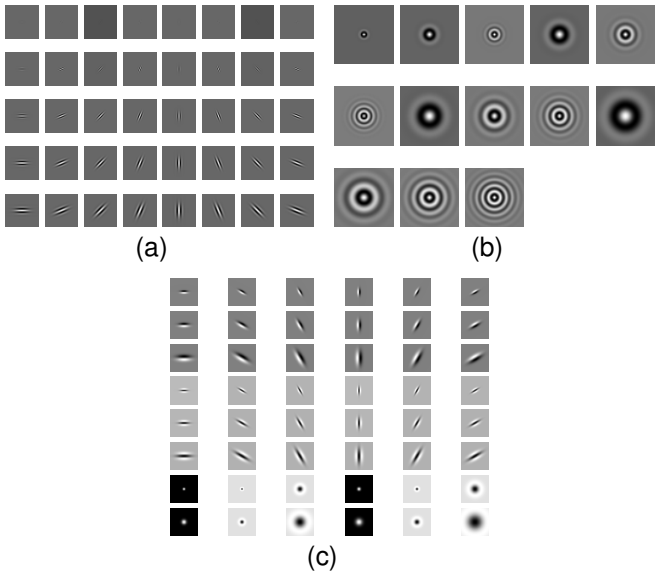

Fig. 3. (a) Gabor filter bank with 8 orientations and 5 sizes; (b) Standard Schimid filter bank; (c) Leung-Malik filter bank. The set consists of first and second derivatives of Gaussians at 6 orientations and 3 scales making a total of 36; 8 Laplacian of Gaussian filters; and 4 Gaussians.

We use different bin quantizations n_c, such that the lightness component of the color space has a coarse representation.

Texture As for the color features, we use texture features to capture the appearance of a person. To deal with object scale and rotation variations, we consider texture features that have invariant properties with respect to these issues. We used a bank of Gabor filters with different sizes and orientations (see Figure 3(a)). After convolving each image with a single filter we computed the modulus of the response and we quantized it in a histogram with g bins. We denote the set of all such histograms as $\{\Gamma_i\}_{i=1}^{I}$, where i indicates the i^{th} Gabor filter. Similarly we used the Schmid filters (Figure 3(b)) to get the set of histograms $\{\Psi_j\}_{j=1}^{J}$, each of which has s bins. Finally we convolve each given image with the Leung-Malik (LM) filter bank consisting of first and second derivatives of Gaussians at 6 orientations and 3 scales, 8 Laplacian of Gaussian (LoG) filters, and 4 Gaussians (Figure 3(c)). After convolving the image with a single filter we quantized the response in a histogram with m bins. $\{\Lambda_k\}_{k=1}^{K}$ is the set of all such histograms, where k indicates the k^{th} LM filter. An example of the responses of the different filter banks is shown in Figure 4.

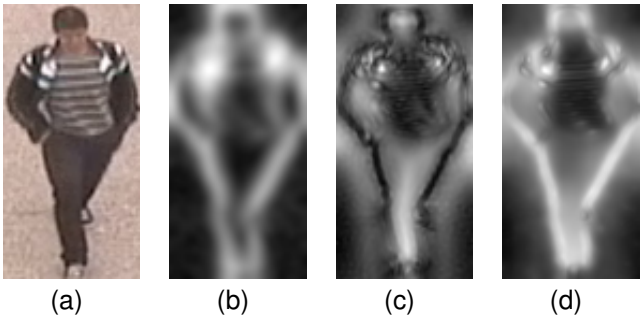

Fig. 4. Response images after convolutions with the three different filter banks shown in Figure 3. All filter responses are sum and scaled for visualization. (a) Input image. (b) Response after convolution of 40 Gabor filters. (c) Response after convolution of 13 Schmid filters. (d) Response after convolution of 48 Leaung-Malik filters.

B. Distance feature vector

Intuition. In the image representation discussed in Section III-A, we compute color, shape and texture features resulting in a very high-dimensional feature vector for each image. Using such a large number of features is advantageous because they can provide a richer representation and capture more subtle visual distinctions between different persons. However, the feature vector may contain non-discriminative elements (some features might capture uninformative features). Even though some invariant properties hold, projecting the feature vector to the feature space of a different camera and match features through proper distances is not always sufficient for finding a good correspondence between persons images. Therefore, we need to find a better way to use the invariant properties of such features and to find the most discriminating elements of the feature vector that allows us to perform a robust re-identification. Towards this objective we propose not to use the distance metrics to find direct correspondences between persons across cameras, but we used the pairwise distance between feature vectors as a new feature.

Distances. To form the DFV for a pair of images we compute pairwise distances for all the considered features. Given two images A and B and the corresponding features extracted as described in Section III-A, we define the following pairwise distances.

- PHOG: $d_\Phi(A^\Phi, B^\Phi)$, where A^Φ and B^Φ are the PHOG features for the image A and image B respectively.
- Color: histograms are compared using distances between feature vectors extracted from the same body part for for each of the three channels as $d_{\Upsilon_T}(A^{\Upsilon_T}, B^{\Upsilon_T})$ and $d_{\Upsilon_L}(A^{\Upsilon_L}, B^{\Upsilon_L})$.
- Gabor: $d_\Gamma(A^{\Gamma_i}, B^{\Gamma_i})$, for $i = 1, \cdots, I$.
- Schmid: $d_\Psi(A^{\Psi_j}, B^{\Psi_j})$, for $j = 1, \cdots, J$.
- LM filters: $d_\Lambda(A^{\Lambda_k}, B^{\Lambda_k})$, for $k = 1, \cdots, K$.

Notice that here we do not specify any particular distance measure since the algorithm can be used with different metrics.

Algorithm 1: Random Forest for Classification of DFVs

Input : Training DFVs
Output: Trained ensemble of trees
for $n \leftarrow$ **to** N **do**
  Draw a bootstrap sample $\mathbf{Z}^*$ of size S from the training data;
  Grow a random-forest tree T_n to the bootstrapped data, by recursively repeating the following steps for each terminal node of the tree, until the minimum node size s_{min} is reached:
  i. Select m variables at random from the p variables.
  ii. Pick the best variable/split-point among the m.
  iii. Split the node into two child nodes.
end
Output the ensemble of trees $\{T_n\}_{n=1}^{N}$;

Classification. The DFV computed for a pair of images of the same person is considered as a positive sample, while the DFV computed for a pair of images of different persons is a negative sample. We use our novel pairwise image representation to discriminate in the distance feature space training a random forest classifier [3].

Let A and B be two images, all the computed distances are concatenated to form the DFV $V_{A,B} = \langle d_{\Upsilon_T}, d_{\Upsilon_L}, \cdots, d_{\Lambda} \rangle$. Then, the goal of classification is to learn a mapping from the feature space of V, to the label space, $Y = \{-1, +1\}$.

The random forests algorithm builds a large collection of de-correlated trees exploiting the bagging idea, where the objective is to reduce the variance of an estimated prediction function by pooling many noisy but approximately unbiased models. Trees are ideal candidates for bagging as they capture complex interaction structures in the data and have low bias. Also, trees are very noisy, hence they benefit greatly from the pooling procedure. As shown in [3], an average of N i.i.d. random variables, each with variance σ^2, has variance $1/N\sigma^2$. If the variables are simply i.d. (identically distributed, but not necessarily independent) with positive pairwise correlation ρ, the variance of the average is $\rho\sigma^2 + \frac{1-\rho}{N}\sigma^2$. As N increases, the second term disappears, but the first remains, and hence the size of the correlation of pairs of bagged trees limits the benefits of pooling. The idea in random forests is to improve the variance reduction of bagging by reducing the correlation between the trees, without increasing the variance too much. This is achieved in the tree-growing process through random selection of the input variables.

To learn the parameters of the decision surface that separates positive and negative DFVs we trained a random forest classifier using the steps given in Algorithm 1. Once the model has been trained, a new sample DFVs $V_{A,B}$ is assigned a class label $\hat{C}^N(V_{A,B}) = majority\,vote\{\hat{C}_n(V_{A,B})\}_{n=1}^{N}$ where $\hat{C}_n(V_{A,B}) = \{-1, +1\}$ is the class prediction of the n-th random-forest tree.

IV. Experimental results

We evaluate the performance our method using two publicly available benchmark datasets: CAVIAR4REID [4] and Wide Area Re-Identification Dataset (WARD) [12]. To show the achieved performance, we computed the Cumulative Matching Characteristic (CMC) curve.

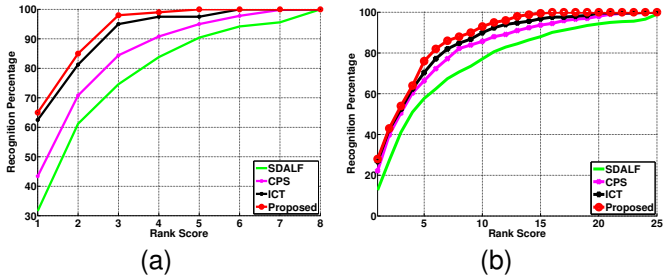

Fig. 5. Comparison of the proposed algorithm with state-of-the-art methods for person re-identification on CAVIAR dataset. In (a) 42 persons have been used for training and 8 person for testing. In (b) 25 persons have been used for training and 25 person for testing.

Implementation details In our current framework, we selected the following settings for all the experiments using 4-fold cross validation.
Shape: PHOG features are extracted for $L = 4$ levels of the spatial pyramid; the HOG histograms computed for each cell have been quantized into $b = 9$ bins.
Color: the histograms for the torso and legs body parts have been computed using 20, 30, and 30 bins for the L*, a*, and b* channel respectively.
Texture: we used Gabor filters at 8 orientations and 5 scales. We used the 13 standard Schmid filters and for LM filters we considered the following. The four basic Gaussians have scales $\sigma = \{\sqrt{2}, 2, 2\sqrt{2}, 4\}$. The first and second derivatives of Gaussians occur at the first three scales with an elongation factor of 3. Finally, the 8 Laplacian of Gaussian filters have been defined using the same σ and 3σ.
Distances: we used the χ^2 distance to compute all the distances given in Section III-B.
Datasets: we followed a standard image normalization procedure on the datasets and we re-sized all the images to 64×128. We tested the performance of our approach using 40 positive and 40 negative samples per person.

A. CAVIAR Dataset

The CAVIAR4REID dataset has 1220 images of 72 persons out of which 50 are acquired by two non-overlapping cameras. The images are of different sizes, varying from 39×17 to 144×72 with illumination and pose changes. We compare our results with those achieved by SDALF [2], CPS [4] and ICT [1], as reported in [1]. To fairly evaluate our approach we used the same two setups proposed in [1] showing the relative performance as a function of the size of the training data. We run 10 independent trials for each setup and average the achieved results.

Figure 5(a) shows the performance of the method when 42 persons have been used to form the training set. The remaining 8 persons form the test set. Using 42 persons as training data our approach achieves similar performance to ICT and it outperforms both the other two methods used for comparison. We achieved 65% rank 1 correct matches and we re-identify all the persons in the test set in the first 5 ranks thus outperforming all other methods.

In figure 5(b), the recognition performance are computed using a training set and a test set of 25 persons. As for the previous scenario, the performances of the our approach are similar to those of ICT. We achieved a recognition percentage of about 78% for a rank score of 5. For the same rank score, a recognition percentage of 72%, 67% and 56% is achieved by ICT, CPS and SDALF, respectively. Similarly as before, we achieve the 100% recognition percentage with a lower rank score value than all other methods. In particular, we recognize all the persons in the test set when the rank score is 15.

B. WARD Dataset

The WARD dataset has 4786 images of 70 persons captured by three non-overlapping cameras. The images are of different sizes, with strong illumination changes. We compare our results with those achieved by RWCAN *et al.* [12] and SDALF [2], then we deeply investigate our performance under two different setups. For each result we run 10 independent trials and we show the average performance for the three camera pairs, here denoted camera pair 1-2, 1-3 and 2-3.

Figure 6(a), 6(b) and 6(c) show the performance of our method compared to RWACN and SDALF. The recognition performance are computed using a training set and a test set of 35 persons. For all the three camera pairs we outperform the methods used for comparisons. For camera pair 1-2 (see Fig. 6(a)), we achieve a recognition percentage of 84% for a rank score of 5, while, for the same rank score, RWACN and SDALF achieve a recognition percentage of 48% and 36% respectively. Similarly, for camera pair 1-3 (see Fig. 6(b)), a recognition percentage of 86% is achieved for a rank score of 5. The other state-of-the-art methods achieve the same recognition percentage for a rank score of 19 and 23 respectively. Finally, for camera pair 2-3 (see Fig. 6(c)), we achieve a recognition percentage of more than 50% for a rank score of 1 thus outperforming both methods used for comparison.

Figure 7(a), 7(b) and 7(c) show the relative performance of the approach as a function of the size of the training data. The CMC curves have been computed using all the color, shape and texture features as described in section III-A. Notice that the maximum rank for each curve is given by the number of persons used for testing. For all the three curves the performance are not decreasing that much even if only 50% of the persons in the dataset are used for training and the remaining 50% of persons for testing. In such case, the worst performances still lead to a recognition rate higher than 33% for the rank 1 score. The best recognition percentage is achieved for the camera pair 2-3, where a recognition rate of 52% is achieved. For all the three camera pairs, the recognition rate strongly improves as the number of persons used for training increases. When 63 persons out of 70 are used for training a recognition rate higher than 60% is achieved for rank 1 for all the camera pairs. In particular, a recognition rate of 81% is reached for rank 1 score for camera pair 1-2.

Figure 8(a), 8(b) and 8(c) show the performance of the method when 56 persons out of 70 have been used to form the training for all the three cameras in the dataset. We show different CMC curves for the remaining 14 persons when only some of the proposed features are used for re-identification. For all the three cameras the combination of all the proposed features achieves the best overall performances. Despite of that it's worth noticing some facts. For the first and the third camera pair, the most discriminative features are the color and the shape features, while for the second camera pair the color features have weaker performance than texture and shape features. Considering the combination of all features we achieved about 50% rank 1 correct matches for the first and second camera pair. For the third camera pair the performance increase significantly and a 70% rank 1 is achieved.

V. Conclusions

In this work we presented a novel approach for person re-identification in a non-overlapping multi-camera scenario. We introduced a method that models not the way features are transformed across camera, but exploits invariant features and robust distance metrics to model how the distances between such features are transformed across non-overlapping cameras. Towards this objective we extracted feature vectors from pairs of persons images viewed in different cameras, and we computed the distance between them to form the DFV. We trained a binary classifier to discriminate between DFVs and to perform the re-identification. To validate the proposed method we compared the performance of our approach to state-of-the-art methods using two publicly available benchmark datasets.

As future works we'll evaluate the proposed algorithm using different global and local features. This, combined with the different distances that can be used to compute the DFVs will give more details about the performance of our approach.

References

[1] T. Avraham, I. Gurvich, M. Lindenbaum, and S. Markovitch. Learning Implicit Transfer for Person Re-identification. In *European Conference on Computer Vision, Workshops and Demonstrations*, volume 7583 of *Lecture Notes in Computer Science*, pages 381–390, Florence, Italy, 2012.

[2] L. Bazzani, M. Cristani, and V. Murino. Symmetry-driven accumulation of local features for human characterization and re-identification. *Computer Vision and Image Understanding*, 117(2):130–144, Nov. 2012.

[3] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, 2001.

[4] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom Pictorial Structures for Re-identification. In *Procedings of the British Machine Vision Conference*, pages 68.1–68.11, 2011.

[5] M. Dikmen, E. Akbas, T. S. Huang, and N. Ahuja. Pedestrian Recognition with a Learned Metric. In *Asian conference on Computer vision*, pages 501–512, 2011.

[6] G. Doretto, T. Sebastian, P. Tu, and J. Rittscher. Appearance-based person reidentification in camera networks: problem overview and current approaches. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–25, Jan. 2011.

[7] M. Eichner, M. Marin-Jimenez, a. Zisserman, and V. Ferrari. 2D Articulated Human Pose Estimation and Retrieval in (Almost) Unconstrained Still Images. *International Journal of Computer Vision*, 99(2):190–214, Mar. 2012.

[8] M. Hirzer, P. M. Roth, K. Martin, and H. Bischof. Relaxed Pairwise Learned Metric for Person Re-identification. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, editors, *European Conference Computer Vision*, volume 7577 of *Lecture Notes in Computer Science*, pages 780–793, Berlin, Heidelberg, 2012.

[9] O. Javed, K. Shafique, Z. Rasheed, and M. Shah. Modeling inter-camera spacetime and appearance relationships for tracking across non-overlapping views. *Computer Vision and Image Understanding*, 109(2):146–162, Feb. 2008.

[10] I. Kviatkovsky, A. Adam, and E. Rivlin. Color Invariants for Person Re-Identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1622–1634, 2013.

[11] C. Liu, S. Gong, C. C. Loy, and X. Lin. Person Re-identification : What Features Are Important ? In *European Conference on Computer Vision, Workshops and Demonstrations*, pages 391–401, Florence, Italy, 2012.

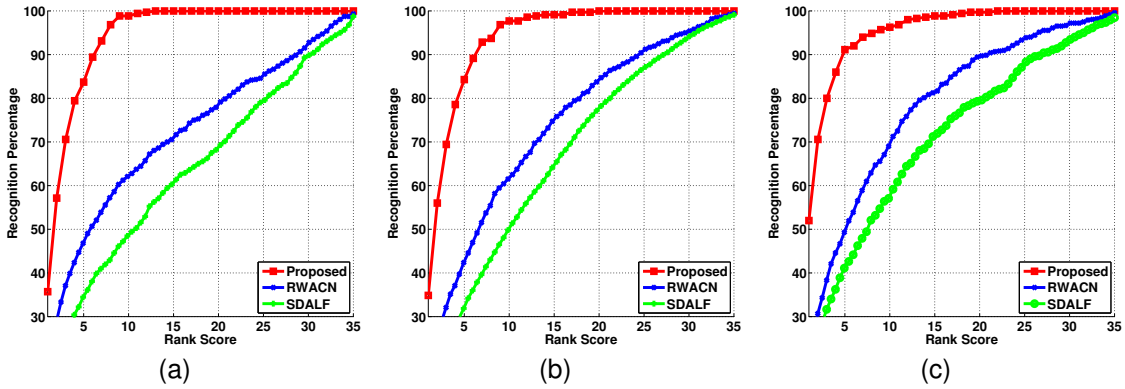

Fig. 6. Comparison of the proposed algorithm with state-of-the-art methods for person re-identification on WARD dataset. (a) Recognition performance for camera pair 1-2. (b) Recognition performance for camera pair 1-3. (c) Recognition performance for camera pair 2-3.

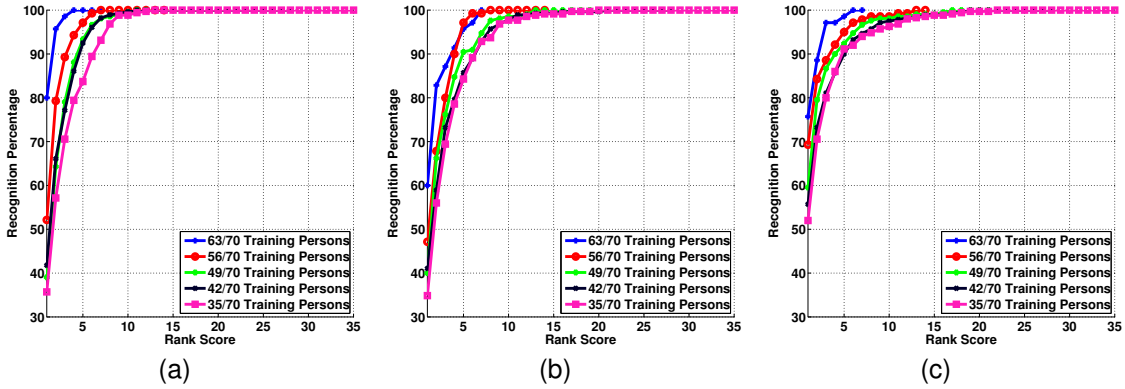

Fig. 7. Performance on the WARD dataset for varying train and test dataset sizes. Recognition performance for camera pairs 1-2, 1-3 and 2-3 are shown in (a), (b) and (c) respectively.

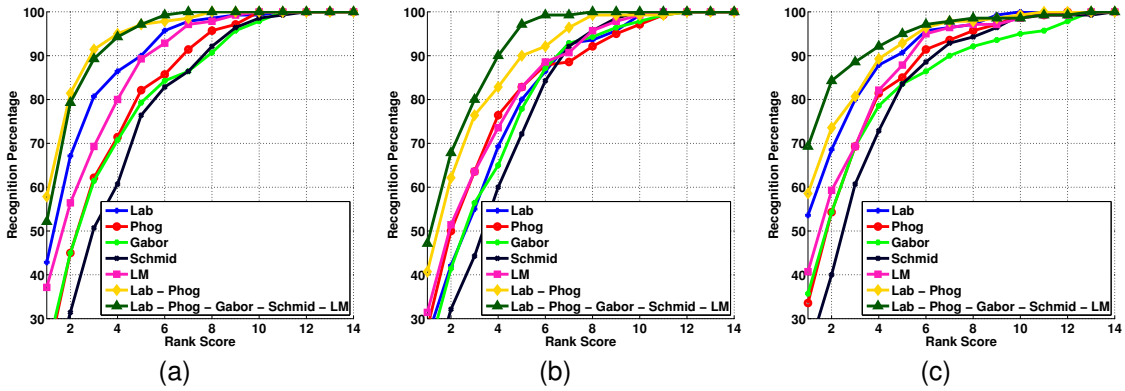

Fig. 8. Performance on the WARD dataset using different combination of the proposed features. Recognition performance for camera pairs 1-2, 1-3 and 2-3 are shown in (a), (b) and (c) respectively.

[12] N. Martinel and C. Micheloni. Re-identify people in wide area camera network. In *International Conference on Computer Vision and Pattern Recognition Workshops*, pages 31–36, Providence, RI, June 2012. IEEE.

[13] N. Martinel, C. Micheloni, and C. Piciarelli. Distributed Signature Fusion for Person Re-Identification. In *International conference on Distributed Smart Cameras*, pages 1–6, Hong Kong, Hong Kong, 2012.

[14] F. Porikli and M. Hill. Inter-Camera Color Calibration Using Cross-Correlation Model Function. In *IEEE International Conference on Image Processing*, pages 133–136, 2003.

[15] Y. Wu, M. Minoh, M. Mukunoki, W. Li, and S. Lao. Collaborative Sparse Approximation for Multiple-Shot Across-Camera Person Re-identification. In *Advanced Video and Signal-Based Surveillance*, pages 209–214, Sept. 2012.

[16] G. Zhang, Y. Wang, J. Kato, T. Marutani, and M. Kenji. Local distance comparison for multiple-shot people re-identification. In K. M. Lee, Y. Matsushita, J. M. Rehg, and Z. Hu, editors, *Asian conference on Computer Vision*, volume 7726 of *Lecture Notes in Computer Science*, pages 677–690, Berlin, Heidelberg, 2013.

[17] R. Zhao, W. Ouyang, and X. Wang. Unsupervised Salience Learning for Person Re-identification. In *International Conference on Computer Vision and Pattern Recognition*, 2013.

[18] W.-S. Zheng, S. Gong, and T. Xiang. Re-identification by Relative Distance Comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):653–668, June 2013.